Received: 5 August 2020 | Revised: 30 September 2020 | Accepted: 17 October 2020

DOI: 10.1111/tbed.13887

SHORT COMMUNICATION

Transboundary and Emerging Diseases WILEY

Google Trends Data and COVID-19 in Europe: Correlations and model enhancement are European wide

Mihály Sulyok[1,2] | Tamás Ferenci[3,4] | Mark Walker[5]

[1]Institute of Tropical Medicine, Eberhard Karls University, Tübingen, Germany

[2]Department of Pathology and Neuropathology, Eberhard Karls University of Tübingen, Tübingen, Germany

[3]Physiological Controls Research Center, Óbuda University, Budapest, Hungary

[4]Department of Statistics, Corvinus University of Budapest, Budapest, Hungary

[5]Department of the Natural and Built Environment, Sheffield Hallam University, Sheffield, UK

Correspondence
Mihály Sulyok, Institute of Tropical Medicine, Eberhard Karls University, Wilhelmstraße 27, Tübingen 72074, Germany.
Email: mihaly.sulyok@uni-tuebingen.de

Summary

The current COVID-19 pandemic offers a unique opportunity to examine the utility of Internet search data in disease modelling across multiple countries. Most such studies typically examine trends within only a single country, with few going beyond describing the relationship between search data patterns and disease occurrence. Google Trends data (GTD) indicating the volume of Internet searching on 'coronavirus' were obtained for a range of European countries along with corresponding incident case numbers. Significant positive correlations between GTD with incident case numbers occurred across European countries, with the strongest correlations being obtained using contemporaneous data for most countries. GTD was then integrated into a distributed lag model; this improved model quality for both the increasing and decreasing epidemic phases. These results show the utility of Internet search data in disease modelling, with possible implications for cross country analysis.

KEYWORDS

COVID-19, Google Trends, model, SARS-CoV-2, surveillance

1 | INTRODUCTION

The current coronavirus (COVID-19) pandemic is probably the most important public health challenge caused by an infectious disease since the Spanish Flu pandemic. It has also become an Internet phenomenon, leading newsfeeds and trending on news forums globally. Understandably, there is widespread public interest, which is being met by blanket media coverage of an unprecedented nature. The Internet is now the favoured first port of call for those seeking healthcare information (Andreassen et al., 2007; Diaz et al., 2002). Therefore, such digital information is likely to be playing a key role in public communication during the current crisis.

Data generated through Internet searching have long been known to be useful for disease monitoring and surveillance (Anema et al., 2014; Brownstein et al., 2009; Eysenbach, 2011; Mavragani et al., 2018). The growing field of research concerning Internet-based healthcare and disease information sources is now widely known as Infodemiology (Eysenbach, 2011). This can be defined as the science relating to the distribution, and factors affecting the distribution, of information in an electronic format, principally upon the Internet, about healthcare (Eysenbach, 2009). More simply, but not exclusively, it is the analysis and study of how people search using the Internet for health-related information. The field first gained recognition with research that tracked influenza using Internet searching patterns (Eysenbach, 2006). Use of online digital data sources offers the potential to enhance not only disease surveillance and monitoring, but could also prove invaluable in disease forecasting and modelling (Reviewed: Salathé et al., 2012).

A favoured source of information for such data is Google Trends, a website which provides data on the volume of Internet searching

upon specific topics. This has been identified as a potentially useful source of real-time data (Carneiro & Mylonakis, 2009; Nuti et al., 2014). Such information may possibly reflect disease occurrence quicker and more accurately than traditional, but slower, disease monitoring through official channels. Studies examining the relationship between Internet searching and disease occurrence using this and similar sites have become commonplace (Carneiro & Mylonakis, 2009; Mavragani & Ochoa, 2019; Mavragani et al., 2018).

However, typically such studies examine relationships between search behaviour and disease occurrence within only a single national country. Whether patterns between search behaviour and disease incidence occur on a broader geographical scale, across national borders, is seldom studied. This is surprising; infectious diseases are often globally distributed, and do not respect national borders. Also importantly, often studies use mainly non-specific symptom keywords, which can cause a loss of specificity in results. On the other hand, selecting overly specific keywords, such as actual disease names, can result in a decrease in sensitivity, and may also make models vulnerable to media-related `noisy` information (Sulyok et al., 2020). As highlighted in a recent review (Mavragani et al., 2018), although many studies examining Internet search behaviour describe relationships and seek correlations, there is a paucity of studies taking the next step and using such data in disease forecasting and modelling.

Thus, here the aim was not only to examine whether correlations between Google Trends data and COVID-19 cases occurred, but also to utilize such data in modelling; could such data enhance traditionally based models using reported case numbers? Additionally, are models enhanced when examining data from a wider geographical range than a single nation state? COVID-19 is a pan-European problem, with epidemics developing almost simultaneously across many countries. This situation provides a unique opportunity to examine whether such data can enhance modelling across multiple countries, continent wide.

2 | MATERIALS AND METHODS

Data relating to Google Internet searching on the single search term 'coronavirus' were downloaded for a range of European countries from the Google Trends website (https://trends.google.com/trends) (Google, 2020) on the 14 March 2020. GTD indexes the volume of search interest against a benchmark index of 100. Data was collected for several European countries where COVID-19 cases have been confirmed. 'Coronavirus' was selected as a search term due to the ubiquitous use of this name in popular parlance across Europe. 'Coronavirus' is part of the official definition of this condition, 'coronavirus disease 2019 (COVID-19)' (WHO, 2020). Google translate showed that 'coronavirus' was commonly used across the majority of European countries involved in our analyses. Another study has demonstrated that results using this search term on Google Trends are highly correlated with related search terms (Walker & Sulyok, 2020).

GTD provides an indexed figure for the volume of Internet searching within a specific nation on a particular chosen keyword. GTD indexes the volume of search interest against a benchmark of 100. Thus data from Google are a reflection of the searching behaviour of people using Google every day (Google, 2020). Data are anonymized and standardized. Data are normalized, as described by Google (2020).

Data were collected on 15 Mar 2020 for a 51-day period running from 23 Jan 2020 to 13 Mar 2020. This encompassed the initial phases of the outbreak, from the potential threat of COVID-19 being highlighted by WHO in a statement on 30 January 2020 (WHO, 2020). Corresponding incidence data were obtained from the GitHub database (https://github.com/CSSEGISandData/COVID-19) of the Coronavirus COVID-19 Global Cases website, managed by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) homepage, which is being updated daily based on WHO, CDC, ECDC, NHC and DXY and local media reports (Dong et al., 2020). Data for the decreasing phase of the European epidemic were obtained similarly on 27 July 2020 for a 90-day period from 27 Apr to 25 Jul 2020. Spearman rank cross-correlation analyses between incident case number and corresponding Google search volumes were performed using a +-40 days lag.

Time series modelling of incident cases was performed using generalized additive models (GAM). The smoothed numerical date was added to the model as an independent variable with and without distributed lag country-specific GTD values. Splines are piecewise defined polynomials which have the ability to smoothly follow non-linear—or even non-monotone data, requiring relatively few degrees of freedom. (Wood, 2004, 2017; Wood & Augustin, 2002). Additional information about the applied spline-based methods is available from (Gasparrini, 2011; Wood, 2004).

Incident case numbers were added as the explained variable. Models were compared with the Akaike Information Criteria. All analyses were performed with R (version 3.6.3) (The R Core Team, 2020) using the mcgv (Wood, 2004) and dlnm package (Gasparrini, 2011).

3 | RESULTS AND DISCUSSION

3.1 | Cross-correlation analyses

Table 1 shows results of Spearman rank contemporaneous correlations between GTD and incident case numbers for different European countries. The increasing and decreasing phases of each country's epidemics were also examined. For the increasing phase, the correlation was positive and significant in all countries studied. With cross-correlation analyses, in the majority of cases the correlation was strongest at a median of 0 day (IQR: 0–1) lag (Figure 1). In other words, GTD was contemporaneous with incident cases (with the notable exception of Ireland, where it preceded it by 18 days).

This pattern was also observed when data describing the decreasing phase of the epidemic were examined. All but two countries had significant moderate to strong positive correlations between GTD and incident case numbers (the exception being Sweden and France). With cross-correlation analyses, the median lag was also 0 days (IQR: −2.0–0, further details in the Supplementary Material S1).

TABLE 1 Spearman rank contemporaneous correlations between GTD and COVID-19 incident daily case numbers for a number of European countries

	Increasing phase		Decreasing phase	
Country	ρ-value	*P*-value	ρ-value	*P*-value
Belgium	0.688	<.001	0.768	<.001
France	0.791	<.001	0.012	.904
Germany	0.808	<.001	0.428	<.001
Hungary	0.470	<.001	0.640	<.001
Ireland	0.405	<.001	0.664	<.001
Italy	0.802	.003	0.843	<.001
Netherlands	0.688	<.001	0.763	<.001
Norway	0.767	<.001	0.479	<.001
Spain	0.779	<.001	0.513	<.001
Sweden	0.805	<.001	0.107	.314
Switzerland	0.716	<.001	0.371	.003
United Kingdom	0.775	<.001	0.804	<.001

Abbreviation: GTD, Google Trends Data.

3.2 | Modelling incident case numbers

Extending standard models, which use solely spline described numerical date as covariate, with GTD improved model quality (AIC without GTD: 2,120.67; with GTD: 2084.72). To test our results, the same model fitting with data describing mostly the decreasing phase of the epidemic was performed. These findings may confirm the results of the previous modelling; model quality was slightly better with the addition of GTD data (AIC without GTD: 10,852.21; with GTD: 10,840.81). Detailed results are available as Supplementary Material S1.

4 | DISCUSSION

Here, results of cross-correlations showed a clear relationship between GTD and reported case incidence across a number of European countries. The quality of time series modelling, as indicated by AIC values, was enhanced by the addition of GTD. Importantly, this enhancement was seen across a number of European countries. This suggests that such data could be of real utility in disease modelling and possibly forecasting in the future, and also that such data could be of value when examining epidemics across country boundaries. This could be of potential utility where traditional disease surveillance is challenging. With the demonstration that GTD data can enhance GAM modelling, more detailed model comparison and validation, with subsequent prediction making, would be the logical next step.

Country-specific factors, possibly reflecting differences in testing and case reporting probably play a critical role. Reported case numbers may not truly reflect disease occurrence, possibly only how vigorous testing regimes are. This was mitigated by examining increasing and decreasing phases of the epidemic separately; reported case numbers are likely to be more reliable at the beginning of an epidemic when the majority of cases can be identified.

Of note, examination of splines for Spain and Italy is notably different in appearance than those for the other European countries studied (S1). This may be related to the timing and rate of development of the COVID-19 epidemic in these countries. Italy experienced a substantial and sustained COVID-19 epidemic prior to other European countries, closely followed by Spain. Other European countries which subsequently experienced epidemics were able to implement measures following the Italian and Spanish lead, which may have affected the pattern of epidemic development in these countries. Detailed model characteristics are provided in the Supplementary Material S1.

Google Trends has proved fertile ground for those wishing to research popular interest in disease or health-related topics. Much research has centred on examining infectious conditions, such as influenza (Eysenbach, 2006; Lazer et al., 2014), but research has covered a range of conditions ranging from tuberculosis (Frauenfeld et al., 2020) to even vector-borne conditions such as tick-borne Encephalitis (Sulyok et al., 2020). Reviews of such studies can be found in Mavragani et al. (2018) and Nuti et al. (2014). Interest in studying Internet search patterns relating to COVID-19 using data from Google Trends has been high, with studies appearing rapidly (e.g. Effenberger et al., 2020). Studies have examined more than simply the relationship between Internet searching and COVID-19 incidence, with topics studied ranging from interest in vaccination (Paguio et al., 2020), anosmia (Panuganti et al., 2020), and even a study of commonly used COVID-19 synonyms (Rovetta & Bhagavathula, 2020a). However, in line with Google Trends studies examining other conditions, most of these COVID-19-related studies concentrate on examining search behaviours within only single countries (e.g. (Husain et al., 2020; Husnayain et al., 2020; Rovetta & Bhagavathula, 2020b; Walker & Sulyok, 2020).

Thus, the examination for countries across Europe as is done here is of particular value. Europe is of particular interest due to the number of countries, some with relatively large population sizes, constrained within a limited geographical area. Many European countries share relatively similar levels of economic development and socio-cultural backgrounds. Movement between European countries occurs at high levels, facilitating infectious disease spread. This allows examination of trends in Internet searching across national borders, when other factors are similar in nature.

This study is also of particular interest given the relative paucity of studies using GTD in modelling. An initial aim was to demonstrate that the addition of GTD to a standard model would enhance its accuracy. An additional novelty being the use of distributed lag GAM. With the demonstration that such data have proved of value, the logical next step would be a series of model comparisons and validations to determine which mode of modelling is most effective, ultimately resulting in the use of GTD for prediction making. Given the emphasis of other GTD studies on correlation and relationship seeking (Mavragani et al., 2018), the modelling performed here is

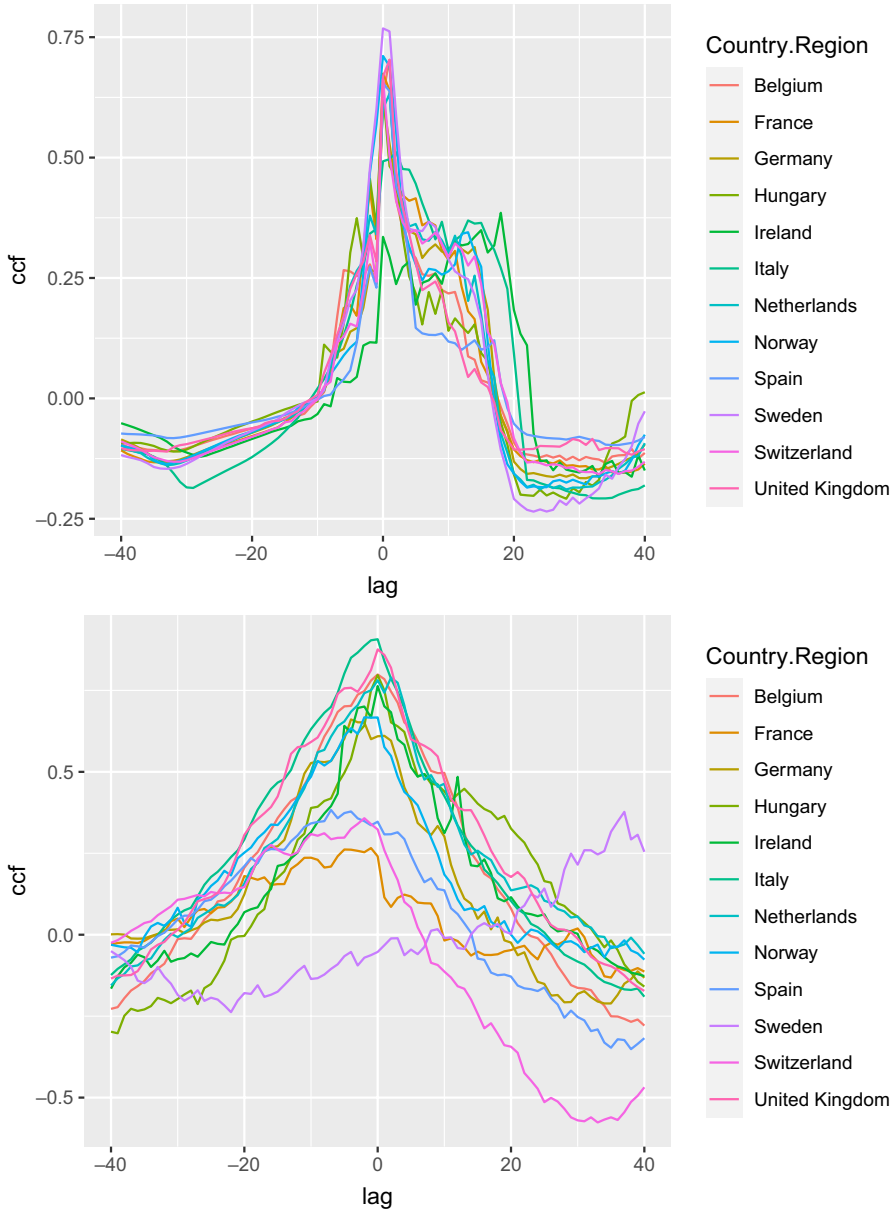

FIGURE 1 Spearman rank Cross-Correlations between GTD and COVID-19 incident daily case numbers for a number of European countries. Upper plot: increasing phase, lower plot decreasing phase of the epidemic (ccf: cross-correlation ρ value) [Colour figure can be viewed at wileyonlinelibrary.com]

therefore a particularly useful addition to the literature and a good start in this direction. The finding that GTD enhanced modelling across a range of European countries is of interest and potentially important.

Although a promising data source, there were initially some problems using Google Trends for prediction purposes. Most notoriously is the use of such data in Google Flu Trends (Lazer et al., 2014), which was dogged by overfitting problems. Increasingly however, refinement of the techniques used means modelling is now more reliable and potentially more useful. Other problems with research using Google Trends data are mainly related to the lack of standard methodologies (Nuti et al., 2014). Use and reporting of the search terms used in analyses, and the method by which searching is performed, often vary considerably between studies (Nuti et al., 2014). Recently, attempts to mitigate against this have been made by establishing common reporting frameworks (Eysenbach, 2009).

In conclusion, GTD showed a strong contemporaneous correlation with incident case numbers across Europe. Patterns between Google Trends data and COVID-19 incidence were found to be of a consistent nature across multiple European countries; an important finding. Using a distributed lag GAM, GTD also enhanced the quality of disease models using solely case numbers for a range of European countries. This improvement suggests such techniques could be used across country boundaries. This is potentially important as COVID-19 reaches new states, especially ones where testing and surveillance are not as reliable as in Europe.

CONFLICT OF INTEREST

We have no conflict of interest to declare.

ETHICAL STATEMENT

The authors confirm that the ethical policies of the journal, as noted on the journal's author guidelines page, have been adhered to. No ethical approval was required since completely anonymized data were obtained from publicly available sources.

DATA AVAILABILITY STATEMENT

All data and the statistical analyses code are available under the link: https://github.com/msulyok/COVID19GoogleTrendsEurope

ORCID

Mihály Sulyok https://orcid.org/0000-0002-6960-5126

REFERENCES

Andreassen, H. K., Bujnowska-Fedak, M. M., Chronaki, C. E., Dumitru, R. C., Pudule, I., Santana, S., Voss, H., & Wynn, R. (2007). European citizens' use of E-health services: A study of seven countries. *BMC Public Health*, *7*, 53. https://doi.org/10.1186/1471-2458-7-53

Anema, A., Kluberg, S., Wilson, K., Hogg, R. S., Khan, K., Hay, S. I., Tatem, A. J., & Brownstein, J. S. (2014). Digital surveillance for enhanced detection and response to outbreaks. *The Lancet Infectious Diseases*, *14*, 1035–1037. https://doi.org/10.1016/S1473-3099(14)70953-3

Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital disease detection–harnessing the Web for public health surveillance. *New England Journal of Medicine*, *360*(2153–2155), 2157. https://doi.org/10.1056/NEJMp0900702

Carneiro, H. A., & Mylonakis, E. (2009). Google trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, *49*, 1557–1564. https://doi.org/10.1086/630200

Diaz, J. A., Griffith, R. A., Ng, J. J., Reinert, S. E., Friedmann, P. D., & Moulton, A. W. (2002). Patients' use of the Internet for medical information. *Journal of General Internal Medicine*, *17*, 180–185. https://doi.org/10.1046/j.1525-1497.2002.10603.x

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*, 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

Effenberger, M., Kronbichler, A., Shin, J. I., Mayer, G., Tilg, H., & Perco, P. (2020). Association of the COVID-19 pandemic with Internet Search Volumes: A Google TrendsTM Analysis. *International Journal of Infectious Diseases*, *95*, 192–197. https://doi.org/10.1016/j.ijid.2020.04.033

Eysenbach, G. (2006). *Infodemiology: tracking flu-related searches on the web for syndromic surveillance*. AMIA Annu Symp Proc 244–248.

Eysenbach, G. (2009). Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *Journal of Medical Internet Research*, *11*(1), e11. https://doi.org/10.2196/jmir.1157

Eysenbach, G. (2011). Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. *American Journal of Preventive Medicine*, *40*, S154–158. https://doi.org/10.1016/j.amepre.2011.02.006

Frauenfeld, L., Nann, D., Sulyok, Z., Feng, Y.-S., & Sulyok, M. (2020). Forecasting tuberculosis using diabetes-related google trends data. *Pathog Glob Health*, *114*, 236–241. https://doi.org/10.1080/20477724.2020.1767854

Gasparrini, A. (2011). Distributed Lag Linear and Non-Linear Models in R: The Package dlnm. *Journal of Statistical Software*, *43*, 1–20.

Google (2020). *Google: Google Trends* [Online] Available at https://trends.google.com/trends/?geo=US (accessed April 6 and July 27, 2020).

Google (2020). *Google: FAQ about Google Trends data - Trends Help* [Online] Available at https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052 (accessed September 28, 2020).

Husain, I., Briggs, B., Lefebvre, C., Cline, D. M., Stopyra, J. P., O'Brien, M. C., Vaithi, R., Gilmore, S., & Countryman, C. (2020). Fluctuation of Public Interest in COVID-19 in the United States: Retrospective Analysis of Google Trends Search Data. *JMIR Public Health and Surveillance*, *6*(3), e19969. https://doi.org/10.2196/19969

Husnayain, A., Fuad, A., & Su, E.-C.-Y. (2020). Applications of Google Search Trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan. *International Journal of Infectious Diseases*, *95*, 221–223. https://doi.org/10.1016/j.ijid.2020.03.021

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, *343*, 1203–1205. https://doi.org/10.1126/science.1248506

Mavragani, A., & Ochoa, G. (2019). Google Trends in Infodemiology and Infoveillance: Methodology Framework. *JMIR Public Health and Surveillance*, *5*, e13439. https://doi.org/10.2196/13439

Mavragani, A., Ochoa, G., & Tsagarakis, K. P. (2018). Assessing the methods, tools, and statistical approaches in google trends research: systematic review. *Journal of Medical Internet Research*, *20*, e270. https://doi.org/10.2196/jmir.9366

Nuti, S. V., Wayda, B., Ranasinghe, I., Wang, S., Dreyer, R. P., Chen, S. I., & Murugiah, K. (2014). The use of google trends in health care research: A systematic review. *PLoS One*, *9*, e109583. https://doi.org/10.1371/journal.pone.0109583

Paguio, J. A., Yao, J. S., & Dee, E. C. (2020). Silver lining of COVID-19: Heightened global interest in pneumococcal and influenza vaccines, an infodemiology study. *Vaccine*, *38*, 5430–5435. https://doi.org/10.1016/j.vaccine.2020.06.069

Panuganti, B. A., Jafari, A., MacDonald, B., & DeConde, A. S. (2020). Predicting COVID-19 Incidence Using Anosmia and Other COVID-19 Symptomatology: Preliminary Analysis Using Google and Twitter. *Otolaryngology - Head and Neck Surgery*, *163*, 491–497. https://doi.org/10.1177/0194599820932128

Rovetta, A., & Bhagavathula, A. S. (2020a). Global Infodemiology of COVID-19: Analysis of Google Web Searches and Instagram Hashtags. *Journal of Medical Internet Research*, *22*, e20673. https://doi.org/10.2196/20673

Rovetta, A., & Bhagavathula, A. S. (2020b). COVID-19-Related Web Search Behaviors and Infodemic Attitudes in Italy: Infodemiological Study. *JMIR Public Health and Surveillance*, *6*, e19374. https://doi.org/10.2196/19374

Salathé, M., Bengtsson, L., Bodnar, T. J., Brewer, D. D., Brownstein, J. S., Buckee, C., Campbell, E. M., Cattuto, C., Khandelwal, S., Mabry, P. L., & Vespignani, A. (2012). Digital Epidemiology. *PLoS Computational Biology*, *8*(7), e1002616. https://doi.org/10.1371/journal.pcbi.1002616

Sulyok, M., Richter, H., Sulyok, Z., Kapitány-Fövény, M., & Walker, M. D. (2020). Predicting tick-borne encephalitis using Google Trends. *Ticks and Tick-borne Diseases*, *11*, 101306. https://doi.org/10.1016/j.ttbdis.2019.101306

Walker, M. D., & Sulyok, M. (2020). Online behavioural patterns for Coronavirus disease 2019 (COVID-19) in the United Kingdom. *Epidemiology and Infection*, *148*, e110. https://doi.org/10.1017/S0950268820001193

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*, 673–686. https://doi.org/10.1198/016214504000000980

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*, 2nd ed.. Chapman and Hall/CRC.

Wood, S. N., & Augustin, N. H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, *157*, 157–177. https://doi.org/10.1016/S0304-3800(02)00193-X

World Health Organization: Statement on the second meeting of the International Health Regulations (2005) *Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)* [Online] Available at https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov) (accessed August 3, 2020).

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Sulyok M, Ferenci T, Walker M. Google Trends Data and COVID-19 in Europe: Correlations and model enhancement are European wide. *Transbound Emerg Dis*. 2021;68:2610–2615. https://doi.org/10.1111/tbed.13887